PhaseMP: Robust 3D Pose Estimation via Phase-conditioned Human Motion Prior

Mingyi Shi[1], Sebastian Starke[2], Yuting Ye[2], Taku Komura[1], Jungdam Won[3]

[1]The University of Hong Kong, [2]Meta, [3]Seoul National University

ICCV23 PARIS

Motivation

- Temporal coherence plays an important role to produce realistic human motion.
- Simple interpolation is commonly used to refine motion jitters - but it doesn't work well for long-term and heavy-occluded frames.
- A periodic feature, called Phase, shows big potential to improve motion quality by describing motion in multi-dimensional sinusoidal space.

Contribution

- We propose a novel motion prior based on the phase manifold for synthesizing feasible motions for various downstream tasks.
- A new optimization framework incorporating phase feature energy, which can work robustly for many challenging scenarios where the observation is incomplete or ambiguous in temporal and spatial domains

Formulation

Phase: A multi-dimensional sinusoidal vector

Given a windows of 2 seconds motion data, we take the joint velocity $X_t \in \mathbb{R}^{3 \times J \times N}$ as input, followed by differentiable FFT layer:

$$A_t, B_t, F_t = \text{FFT}(\text{Conv}(X_t)) \tag{1}$$

Then the Phase feature P_t is defined with:

$$P_t = [p_t, F_t, A_t]\,, \quad p_t = (A_t \cdot \sin(2\pi \cdot S_t), A_t \cdot \cos(2\pi \cdot S_t)) \tag{2}$$

Prior: A autoregressive generation network

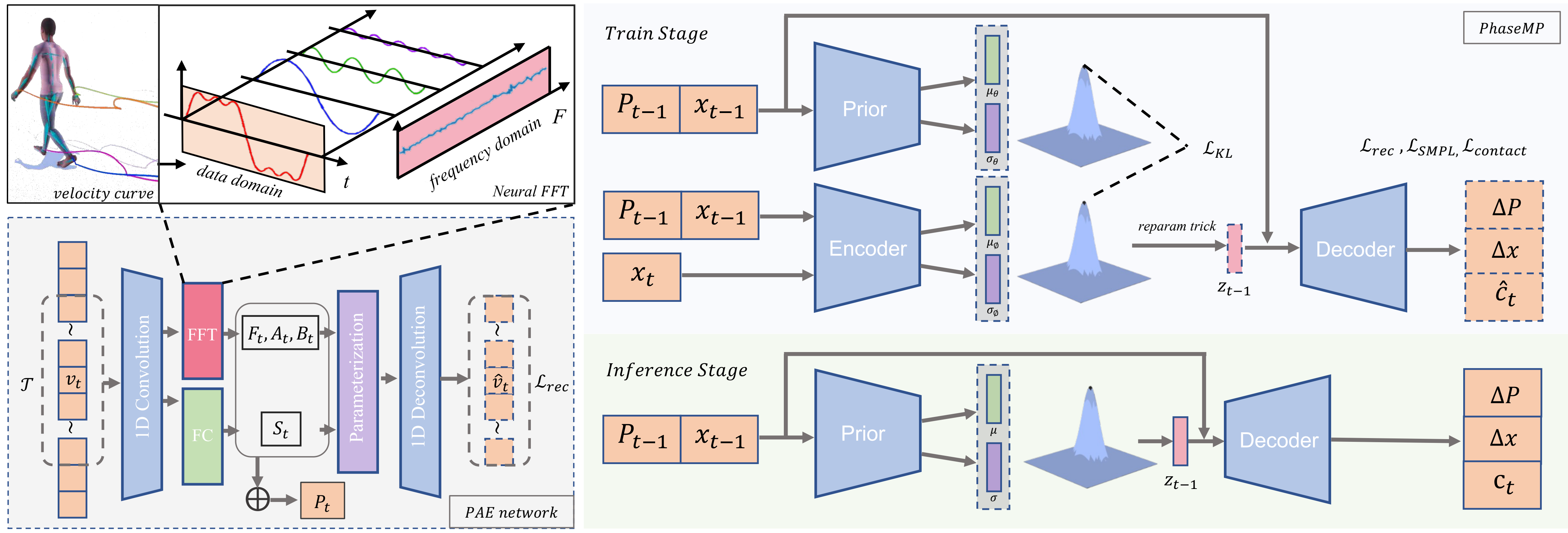
Model the transition between two frames $[x_{t-1}, x_t]$, with condition variant z and phase P

$$\Delta x, \Delta P, c_t = G(x_{t-1}, z_{t-1}, P_{t-1}) \tag{3}$$

$$\hat{x}_t = x_{t-1} \oplus \Delta x, \hat{P}_t = P_{t-1} + \Delta P \tag{4}$$

Then a sequence of poses can be generated.

Framework



Left: Phase feature extraction. The key module is a periodic auto-encoder equipped with convolution and FFT (Fast Fourier Transform) layers in its intermediate structure, allowing it to compute embeddings in the frequency domain given joint velocities as inputs.

Right: Conditional human motion prior. The pose in the next frame is predicted by sampling from a Gaussian distribution produced by the prior model. In the **training** stage, the prior model R is trained with posterior E by aligning their output distributions, without the input of paired frames. Thus, the prior model only predicts solely at **inference** time by only considering the previous frame. Note that we used a sine activation layer in the decoder.

Test-time optimization: Refine the pose sequence

Given partial observations, such as a 2D landmark sequence or a partial 3D joint sequence, we estimate the original 3D pose sequence through optimization.

Stage 1: Produce initial guess of [body pose parameter $\hat{x}_{0:T}$, environment parameters].

Stage 2: Produce the initial [phase feature $P_{1:T}$, variant feature $z_{1:T}$] using the encoder E. Then produce a target phase curve based on our cyclic updating and robust blending strategy.

$$p_t = \bar{A}_t \cdot I(\alpha_P)(R(\theta) \cdot p_{t-1}, (p_{t-1} + \Delta p))\,, \quad \theta = \Delta t \cdot 2\pi \cdot \bar{F}_t \tag{5}$$

Where the $\bar{A}_t$ and $\bar{F}_t$ is dynamic blended based on the confidence value from observation.

Stage 3: Refine the initial data with the following energy:

$$\text{argmin}_{z_{1:T-1}, \beta}(E_{obs} + E_{prior} + E_{reg} + E_{phase}) \tag{6}$$

Experiments

We show the evaluation on following tasks: (i) motion generation (2) motion estimation from sparse observation.

	per-frame reconstruction			sampling (10s)		
Model	Contact ↑	MPJPE↓	PJPE-std↓	Contact ↑	ADE ↓	FDE ↓
HuMoR(MLP, w/o Phase)	0.9770	0.022	0.051	0.8216	45.43	62.47
Ours(MLP, with Phase)	0.9764	0.020	0.040	0.8525	39.48	54.96
Ours(SirenMLP, w/o Phase)	0.9788	0.019	0.031	0.8577	**35.47**	49.95
Ours(SirenMLP, with Phase)	**0.9799**	**0.017**	**0.021**	**0.8662**	42.12	**49.47**

Table 1. Comparison results on AMASS dataset reconstruction

		fitting (3s)						
Method	Input Conditions	Vis	Occ	All	Contact ↑	Accel	P-Frep	P-Dis
VPoser-t	J_{height}>0.9	**0.67**	20.76	9.22	-	5.71	16.77%	2.28
MVAE		2.39	19.15	9.52	-	7.12	3.15%	0.30
HuMoR		1.46	17.40	8.24	**0.89**	5.38	3.31%	**0.26**
Ours		3.94	**15.63**	**8.31**	**0.89**	**4.58**	**3.04%**	0.28
HuMoR	$J_{\text{end\_effectors}}$	**3.05**	4.12	3.83	0.96	4.91	0.31%	1.03
Ours		3.16	**4.07**	**3.79**	**0.97**	**4.88**	**0.28%**	**1.02**
HuMoR	10 frames interval	5.56	7.76	7.49	0.91	7.72	1.57%	1.90
Ours		**3.19**	**4.92**	**4.33**	**0.93**	**6.33**	**1.31%**	**1.72**

Table 2. Comparison results on estimation from different input conditions

Results

HuMoR

Ours